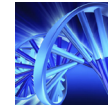


Multiple Sequence Alignment

E - 1



Sequences

> Yeast YOR020c

```
mstllksaksivplmldrnlvqrikaqaktasgylpe
knveklngaeavavpgpfdangnkvpqkvvgdqvl
ipqfggstiklgnddevilfrdaeilakiakd
```

> Neurospora crassa

```
mattvrvskslipldrnlvqrvkaeaktasgflpe
svvkdlnakvlavpggaldkdkrlpmpgvnagdrvl
ipqyggspvkgeeyhlfrdseilakiae
```

> Aspergillus nidulans

```
m5llrnvknlapldrnlvqrvkaeaktasgflpes
svkeqneakvlavpggavdrngqripmgvaagdrvl
pqfggspkigeeeyhlfrdseilakiae
```

> Schizosaccharomyces pombe (fission yeast)

```
matkllksaksivplldrnlvqrikadktasgflpe
ksveklsegsvlsvgkgyngkaglaqpsvavgdrvl
lpaygg5nikvgeey5lyrdhellaiike
```

> Mortierella alpina

```
masritkfskt5vpmmdrnlvqrikpqqktasgiyp
ekaqaalnegyvavgkglttqegkvpselaegdkv
lppyyggsvvkvndeel5fr5eilakiq
```

> Crypthecodinium cohnii

```
matgiakfrfplldrnlvqrlkpeaktasgflpesa
akapnyatvlavpgpgrtdqdlpmnvkvgdkvvp
eygqmtlkfedee5f5vrdadimgilne
```

> Drosophila melanogaster

```
maaaikkipmldrnlvqraealtktkgyvlpekv
gkvlqgtvlavpgpgrtnastgnhipigvgedrvllp
efggtkvnlegdqkelfresdilakle
```

> Homo sapiens

```
aggafrkflplfdrnlvrsaaetvtkggimlpeksq
gkvlqatvavsgsgkkggeiqpsvkvvgdkvlpe
yggtkvvliddkdyflfrdxilgky
```

> Geobacillus stearothermophilus

```
vlkplgdrvvievieteeaktasgvlpdtakekpgqg
rvavvgkrvld5gervapevevgdriifskyagtev
kydgkeylilresdilavig
```

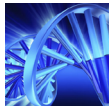
> Mycobacterium tuberculosis

```
makvnikpledilvqaneatt5asgvlpdtakek
pqegtvavagprvrdedgkripld5aegdtviysky
ggt5kyng5eylil5ardv5lav5vk
```

> Mus musculus (house mouse)

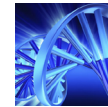
```
maqgafrkflilfdrnlvrsaaetvtkggimlpeks
qgkvlqatvavsgsgkkgseiepsvkvvgdkvlplp
eyggtkvvliddkdyflfrdsdilgkykn
```

E - 2



Multiple Sequence Alignment (MSA)

E - 3



Why MSA?

- Proteins are often related to a larger group (i.e., a family) of proteins
- Multiple sequence alignment is more sensitive than pairwise alignment for detecting homologs
- MSAs can elucidate conserved residues, motifs, or other functional regions in a protein
- MSA is critical for phylogenetic analysis
 - Selection of sequences
 - Multiple sequence alignment of sequences
 - Tree building
 - Tree evaluation

E - 4

```

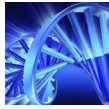
Homo      ----AGQAFRKFLLPLFDRVLVRSAAETVTKGGIMLPEK50GKVLQATVAVVAVGSGK-GK 55
Mus       ----MAGQAFRKFLLPLFDRVLVRSAAETVTKGGIMLPEK50GKVLQATVAVVAVGSGK-GK 56
Drosophila ----MAAAIKKIIPMLDRILIQRAEALTKTKGGIVLPEKAVGKVLQETVLAVGPGTRNAS 56
Neurospora -MATTVRSVKSLIPELLDRVLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 58
Aspergillus -MSLLRN5KNLAPLDRVLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 57
Crypthecodinium ---MATGIAKRFTPLLDRLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 56
Yeast     MSTLL-KSAKSIVPLMDRLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 58
Schizosaccharomyces MATKL-KSAKSIVPLMDRLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 58
Mortierella MASRITKFSKTI5VPMMDRVLVQRVKAETAASGIFLPESSVKDLNEAKVLAVGPGAL-DK 59
Geobacillus -----VLKPLGDRVVI5EVIEETEEKTASGIVLPDTAKEK50PQEGRVVAVGKGRVLD5 50
Mycobacterium -----MAKVNIKPLEDKILVQANEATT5ASGIVLPDTAKEK50PQEGRVVAVGKGRVLD5 54
          : : *::: : . * : * : * . . * : * *

```

```

Homo      GGEIQPVS5KVGDKVLLPEYGGTKVVLDD--DKDYFLFRD5XILGKY--- 99
Mus       SGEIEPVS5KVGDKVLLPEYGGTKVVLDD--DKDYFLFRD5XILGKYVN- 102
Drosophila TGNHPI5GVEKEDRVLLEPFGGTVKNLEGGDKELFLFR5RESILAKLE-- 103
Neurospora DGKRLP5MVAAGDRVLIPQYGGSPVKVGG--EE5YTLFRD5EILAKIAE- 104
Aspergillus NGQRI5PMGVAAGDRVLIPQYGGSPVKVGG--EE5YTLFRD5EILAKIAE- 103
Crypthecodinium DGDIL5PMN5KVGDKVVLPEYGGT5LKE--DEE5FVFRD5ADIM5ILNE- 102
Yeast     NGNKV5PQV5KVGDKVVLPEYGGT5LKE--DEE5FVFRD5ADIM5ILNE- 106
Schizosaccharomyces EGKLA5QPSVAVGDRVLIPQYGGSPVKVGG--EE5YTLFRD5EILAKIAE- 104
Mortierella EGKVP5SELAEGDKVLLPEYGGSPVKVGG--NEE5LILFR5EILAKIQ-- 104
Geobacillus GE-RV5APEVE5GDRIF5SKYAGTE5VKYD--GKE5YTLFR5EILAKIQ-- 94
Mycobacterium GEKRI5PLDV5AEGD5TVI5SKYGGTE5IKYN--GEE5YTLFR5EILAKIQ-- 100
          : ** : : . : * : : . : : :

```

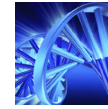


Pairwise Alignment

	A	C	G	T	G	A	
	0	-6	-12	-18	-24	-30	-36
A	-6	5	-1	-7	-13	-19	-25
G	-12	-1	1	4	-2	-8	-14
C	-18	-7	4	-2	0	-6	-12
G	-24	-13	-2	9	3	5	-1
T	-30	-19	-8	3	14	8	2
T	-36	-25	-14	-3	8	10	4
A	-42	-31	-20	-9	2	4	15

AGCGTTA
 | | | |
A-CGTGA

E - 5

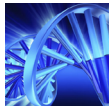


3-Sequence Alignment

AGA
 AGTC
 TCCTA

	A	G	T	C	
T	0	-6	-12	-18	-24
A	-6	5	-1	-7	-13
G	-12	-1	10	4	-2
A	-18	-7	4	6	0

E - 6



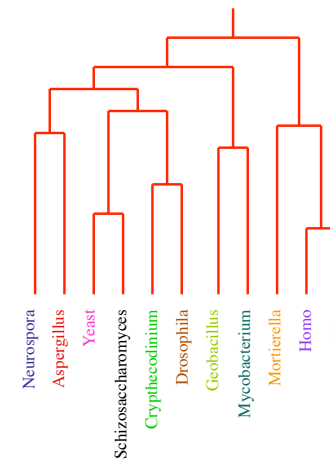
Pairwise Alignment Scores

	Yeast	Neurospora	Aspergillus	Schizosaccharomyces	Mortierella	Cryptocodium	Drosophila	Homo	Geobacillus	Mycobacterium	Mus	
49	46	78	45	55	54	44	38	37	42	Yeast		
	52	41	40	43	46	44	41	39	43	Neurospora		
		43	48	45	45	40	40	38	39	Aspergillus		
			42	53	55	41	41	40	40	Schizosaccharomyces		
				43	46	40	43	38	39	Mortierella		
					61	43	34	36	45	Cryptocodium		
						49	42	36	49	Drosophila		
							37	32	93	Homo		
								59	38	Geobacillus		
									32	Mycobacterium		
										Mus		

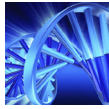
E - 7



Construct a Guide Tree



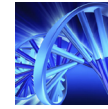
E - 8



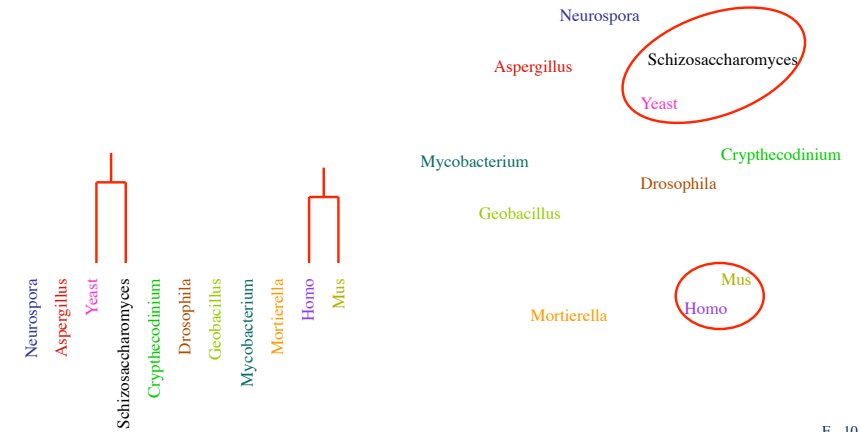
Unweighted Pair Group Method with Arithmetic mean (UPGMA)

- Assume each organism is its own group
- Repeat the following step
 - Merge together the two closest groups

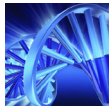
E - 9



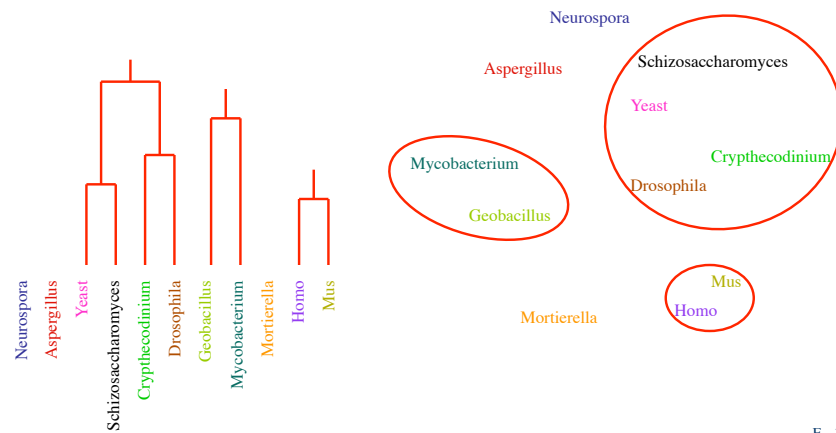
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



E - 10



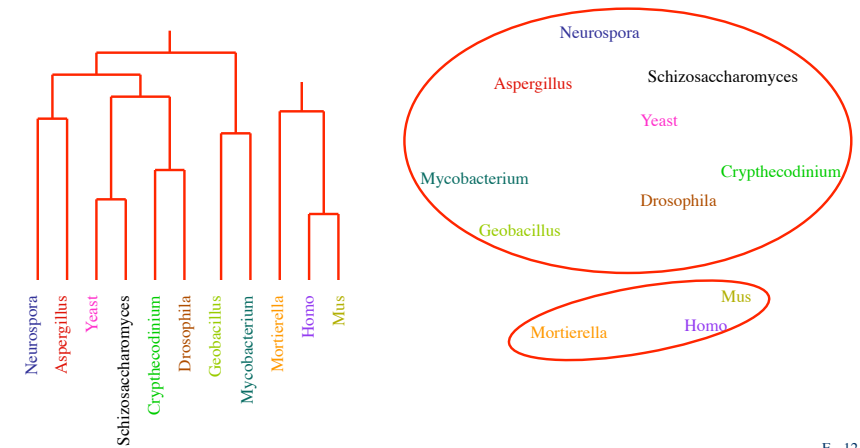
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



E - 11



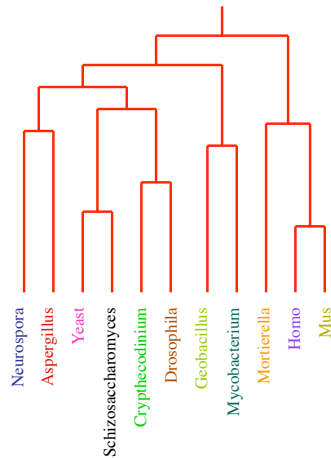
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



E - 12



Guide Tree



E - 13



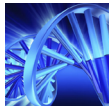
Multiple Sequence Alignment

```

Homo/1-109      ----KQQAFLFLFDVFLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---
Mus/1-109      ---MAGQAFKFLFLFDVFLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---
Drosophila/1-109 ---MAAAKRIIPLDRLIQAEALTFGGIVLPERAVGRLDTVLAVSGGTRNASTENHPIGVKEDRVLLPFGGQFVLE--KDYFLFDGELGK---
Neurospora/1-109 ---MATVRSVSLIPLDRLIQAEALTFGGIVLPERAVGRLDTVLAVSGGTRNASTENHPIGVKEDRVLLPFGGQFVLE--KDYFLFDGELGK---
Aspergillus/1-109 ---MELDNTVLAELLDVFLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---
Cryptocodium/1-109 ---MATEIARFLLDRLVQKFKHATAGLFLPESAAKANIATVAVVSGGTRNASTENHPIGVKEDRVLLPFGGQFVLE--KDYFLFDGELGK---
Yeast/1-109     MSTLI-KSASIVELDRVLDVQKFAATAGLFLPERNVEINAEVAVVSGGTRNASTENHPIGVKEDRVLLPFGGQFVLE--KDYFLFDGELGK---
Schizosaccharomyces/1-109 MATRI-KSASIVELDRVLDVQKFAATAGLFLPERNVEINAEVAVVSGGTRNASTENHPIGVKEDRVLLPFGGQFVLE--KDYFLFDGELGK---
Mortierella/1-109 MNRDRETFEETVYVMSQVFLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---
Geobacillus/1-109 -----VEKPLDQVFLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---
Mycobacterium/1-109 -----MAKVNIFELDRLVLSANITVSGIHLFKSQCVLACTVAVVSSGK-E-GEIQCWFKVGDVLLPFGGQFVLE--KDYFLFDGELGK---

```

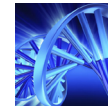
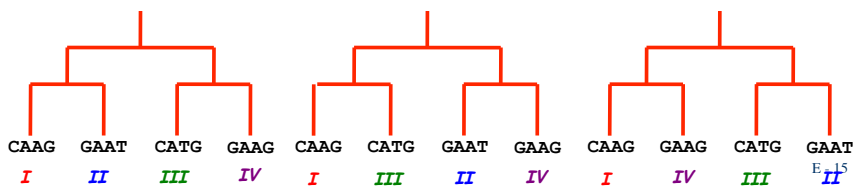
E - 14



Constructing a Phylogenetic Tree

When analyzing a set of data, there are *many* possible phylogenies to consider. We would like to identify a good (the best) phylogeny.

- I** CAAG
- II** GAAT
- III** CATG
- IV** GAAG



Constructing a Phylogenetic Tree

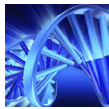
Character based methods are one type of approach for constructing a phylogenetic tree.

Character based methods are often based on the idea of *maximum parsimony*.

" IT IS VAIN TO DO WITH MORE WHAT CAN BE DONE WITH FEWER"
 OR
 Principle of parsimony
 OR
 ...smallest number of evolutionary changes...

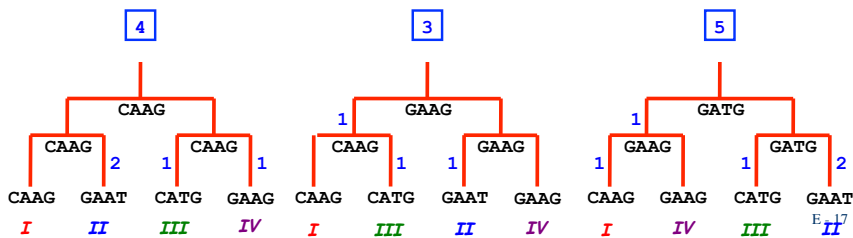
Optimality criterion: The 'most parsimonious' tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences.

E - 16



Character Based Methods

- I** CAAG
- II** GAAT
- III** CATG
- IV** GAAG



Maximum Parsimony Methodology

Step 1: Identify informative sites

Sites with at least two different characters at the site, each of which is represented in at least two of the sequences

	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 18



Maximum Parsimony Methodology

Step 1: Identify informative sites

Sites with at least two different characters at the site, each of which is represented in at least two of the sequences

	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 19

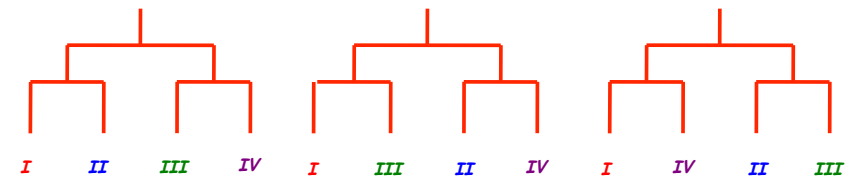


Sites Where All Trees Require the Same Number of Changes Are Not Informative

Tree I

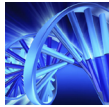
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 20

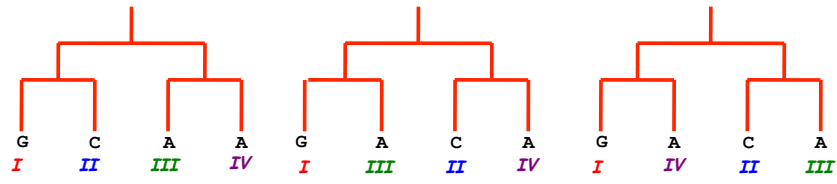


Sites Where All Trees Require the Same Number of Changes Are Not Informative

Tree I

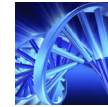
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 21

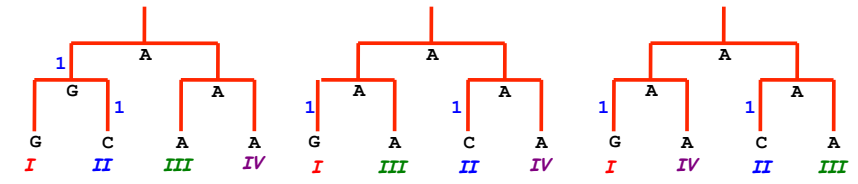


Sites Where All Trees Require the Same Number of Changes Are Not Informative

Tree I

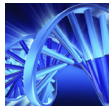
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 22

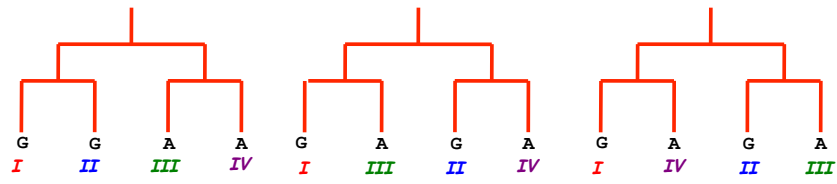


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 23

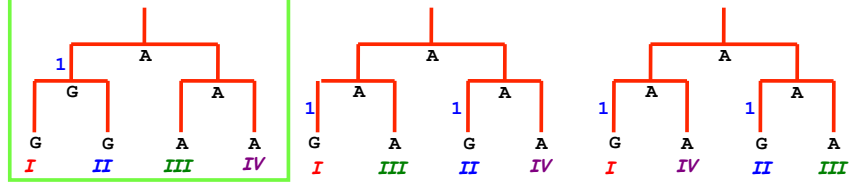


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 24

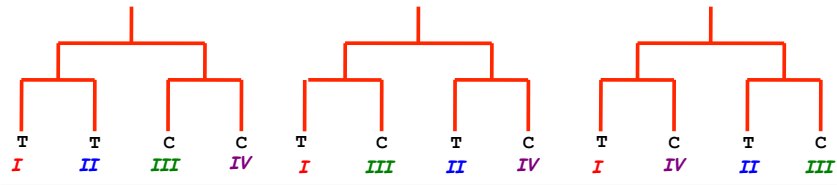


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

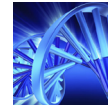
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 25

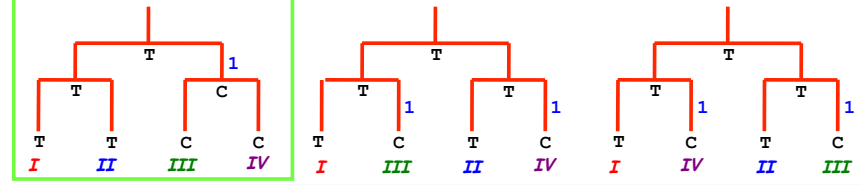


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

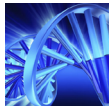
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 26

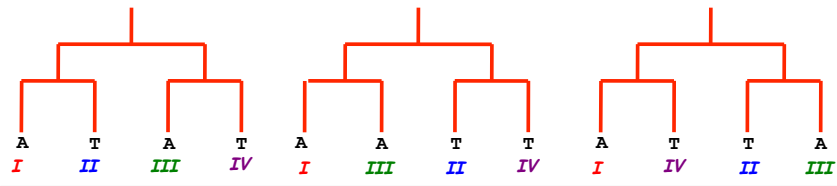


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

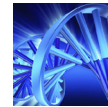
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 27

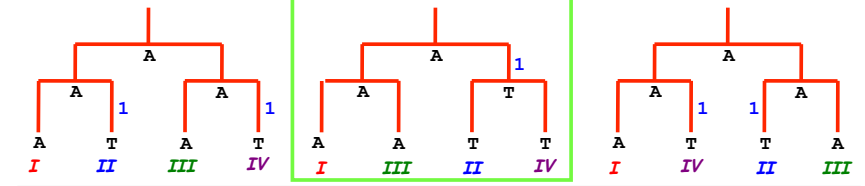


Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes

Tree I

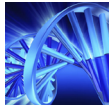
Tree II

Tree III



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

E - 28



Maximum Parsimony Methodology

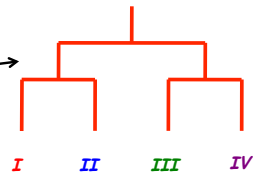
Step 2: Calculate minimum number of substitutions at each informative site

Step 3: Sum number of changes at each informative site for each possible tree

The tree with the least number of total changes is the most parsimonious tree

Number of Changes at Each Informative Site

	5	7	9	Σ
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III	2	2	2	6

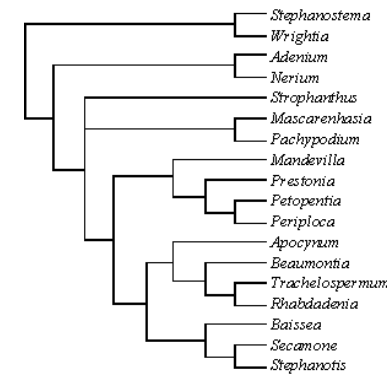


E - 29



How Confident Are We In Our Tree?

Bootstrapping: Given a particular tree, how consistently does a tree-building algorithm find that branching order using a randomly sampled version of the original dataset?



E - 30



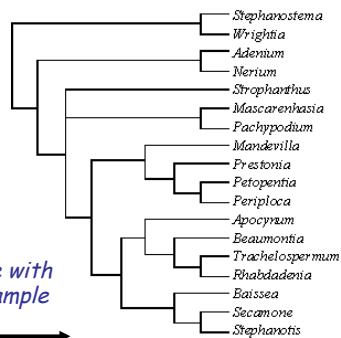
Random Sampling



Resample with replacement



Build tree with pseudosample



E - 31



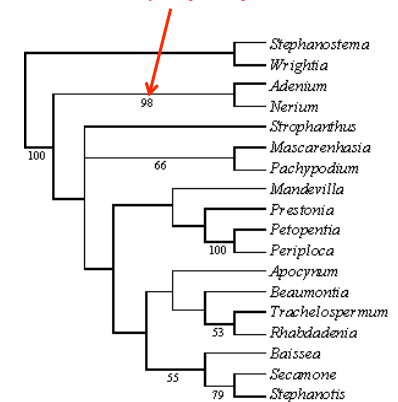
Bootstrapping

How reliable is our tree?

Repeat the following many times:

- Generate a new set of sequences by randomly sampling the original sequences
- Construct a tree for the new set of randomly sampled sequences
- Compare the new tree (based on the random sequences) with the original tree (based on the original sequences) and determine how many times the same tree structures were recovered

Adenium and Nerium were grouped together in 98% of the trees built from randomly sampled sequences



E - 32